Simulation Study of Cascaded On-Off Model

(under the supervision of Prof. J.S. Marron)

Juhyun Park

2/14/01

Department of Statistics
UNC-CH

Motivation

- Internet Traffic Data: observed at an arbitrary point between servers and clients

- How TCP connection works: Once packets are transmitted, the next packets won't be sent until the acknowledgement is arrived.

Show CombineSessionData1p1.pdf

- Expect periodicity

- In case of packet lost, long waiting time appears

Traffic Model

Goal: Find a model for individual TCP Connection traces, that:

1. "Looks qualitatively right"
2. Gives "correct" statistical properties (dependence, …)
3. Aggregates "correctly" (scaling, multifractal, …)
4. Fits easily into queueing analysis.
5. Makes "physical" sense

Cascaded On – Off Model

Ideas:

- each packet is a “rapid burst” (on times)
- waiting times (off times) in between are very diverse (orders of magnitude different)

Mathematical Formulation:

I. Independent On – Off Processes, $X_1(t), X_2(t),...$
where $X_n(t)$ is

"on" for exponential times, with rate $2^{n-1} l$

"off" for exponential times, with rate $2^{n-1} m$

Mathematical Formulation (cont)

II. Vary the “gap distribution” by multiplying:

$$Y_n(t) = \prod_{i=1}^{n} X_i(t)$$

III. Normalize to keep overall expected value the same:

$$Z_n(t) = meanrate\left(\frac{l+m}{m}\right)^n Y_n(t)$$

Note:

The model is characterized by three parameters, l , m and n,

which can be estimated using data later on.

Example of Cascaded On-Off Model

Show lower left of CascOnOff\CascOnOffDemo1.ps

Generate processes with three different rates and multiply them

- Blue: One process
- Magenta: Two processes
- Red: Three processes

Notice:

- The more multiplication occurs, the longer off times appears.
- The longer off times result in the steeper slope in order to send same data/unit time.

Fit Model to Data

Idea: use

- "peak rate" $= r_{peak} = 155*10^6\ (bits/\sec)\big/8\ (bits/byte)$

- N number of packets in trace

- $T_i(t)$ time stamp (secs) of i-th packet

- $S_i(t)$ size (bytes) of i-th packet

to estimate parameters: l, m, n

Parameter Estimation 1

For a given value of the level n

1. “Get total size right”, i.e: est. the “mean rate”, r_{mean}, by

$$\hat{r}_{mean} = \frac{\sum_{i=1}^{n} S_i}{T_N} = \frac{\text{"}Total\ Size\text{"}}{\text{"}Total Time\text{"}}$$

2. “Make jumps right”, i.e: est. the “mean on time”, t_{on}, by

$$\hat{t}_{on} = \frac{\hat{r}_{mean}}{r_{peak}} \cdot \frac{T_N}{N} = \text{"}prop'n\ on\text{"} \cdot \text{"}time / packet\text{"}$$

Parameter Estimation, 2

Still for a given value of the level n

3. "Time conservation" gives the "mean off time", t_{off}, as:

$$\hat{t}_{off} = \frac{T_N}{N} - \hat{t}_{on} = \text{"}time / packet\text{"} - \text{"}mean\ on\ time\text{"}$$

4. Solve rate equations to get:

$$\hat{l}_n = \frac{1}{\hat{t}_{on}\left(2^n - 1\right)}$$

$$\hat{m}_n = \frac{\hat{l}_n}{\left(\hat{t}_{off} / \hat{t}_{on} + 1\right)^{1/n} - 1}$$

5. Estimate n by variance matching

Example of Estimation Process

Show CombineCascOnOffData2p1t1.pdf

1. Start with original trace

2. Compute rate parameters for different n

3. For triple of parameters, find one which gives the closest theoretical variance to sample variance

4. Simulate 5 estimated traces

Results for Cascaded On-Off Model

- Good visual impression

- Statistical summaries?

- Aggregate properties?

- Queueing analysis is tractable

- Physical Sense: delays appear at different levels:
  i. individual packets
  ii. TCP window
  iii. Buffer overflow – packet loss

Again show CascOnOff\CascOnOffDemo1.ps

Simulation of Estimation

Show CombineCascOnOffData3p1n12.ps

Idea: better understand estimation process

1. For real traces, estimate $\hat{n}$, $\hat{l}_{\hat{n}}$ and $\hat{m}_{\hat{n}}$ as above.

Show top of CombineCascOnOffData3p1n12.pdf

2. Simulate 100 traces, as above.

Show bottom right of CombineCascOnOffData3p1n12.pdf

3. Get 100 simulated estimates, $\hat{\hat{n}}$, $\hat{\hat{l}} = \hat{\hat{l}}\left(\hat{\hat{n}}\right)$ and $\hat{\hat{m}} = \hat{\hat{m}}\left(\hat{\hat{n}}\right)$, from simulated traces.

Simulation of Estimation, (Cont.)

Show CombineCascOnOffJHPData3p2n12.ps

A. For computational speed, restricted $n \leq 12$

B. Compare sim'd $\hat{\hat{l}}$ (red), and $\hat{\hat{m}}$ (blue), with "true values" $\hat{l}$ (magenta) and $\hat{m}$ (cyan).

Show CombineCascOnOffData3p2n12.pdf, upper left

i. Sometimes "est's too big" :

Show CombineCascOnOffData3-Big.pdf

ii. Sometimes "est's about right" :

Show CombineCascOnOffData3-OK.pdf

iii. Sometimes "est's too small" :

Show CombineCascOnOffData3-Small.pdf

Simulation of Estimation, (Cont.)

Show CombineCascOnOffData3p2n12.pdf

C. Looks like “clusters” - investigate with “smooth histogram”:

Show CombineCascOnOffData3p2n12.pdf, upper right

-“factor of 2 between peaks”?

D. Joint distributions of $\hat{\hat{l}}$ and $\hat{\hat{m}}$

Show CombineCascOnOffData3p2n12.pdf, lower left

- usually have strong relationship (not surprising)

- don’t lie on a line???

Simulation of Estimation, (Cont.)

E. Joint distribution of $\hat{\hat{t}}_{on}$ and $\hat{\hat{t}}_{off}$:

Show CombineCascOnOffData3p2n12.pdf, lower right

- more "independent" than $\hat{\hat{l}}$ and $\hat{\hat{m}}$.

- So is this a "better parametrization"?

- Simulated $\hat{\hat{t}}_{on}$ always << $\hat{t}_{on}$

- Simulated $\hat{\hat{t}}_{off}$ usually < $\hat{t}_{off}$

Simulation of Estimation, (Cont.)

Show CombineCascOnOffData3p3n12.pdf

F. Investigation of "clustering"

- Clusters explained by $\hat{\hat{n}}$
- Then have $\hat{\hat{n}}$ mostly >> $\hat{n}$
- Otherwise $n \leq 12$ constraint "takes over"???
- Note: "factor of 2" between peaks

Explanation of "Factor of 2"

For $n \to \infty$:

$$\hat{\hat{l}} = \frac{1}{\hat{\hat{t}}_{on}\left(2^n - 1\right)} = \frac{1}{2^n} \cdot \frac{1}{\hat{\hat{t}}_{on}} + o\left(\frac{1}{2^n}\right)$$

$$\hat{\hat{m}} = \frac{\hat{\hat{l}}}{\left(\hat{\hat{t}}_{off} \big/ \hat{\hat{t}}_{on} + 1\right)^{1/n} - 1} = \frac{1}{2^n} \cdot ??? + o\left(\frac{1}{2^n}\right)$$

Should reparametrize, and work with $\hat{\hat{t}}_{on}$ and $\hat{\hat{t}}_{off}$???

Alternative Parameterization:

Show CombineCascOnOffData3p5n12.ps

$$l^* = 2^n \cdot l\ , \qquad m^* = 2^n \cdot m$$

{or $\left(2^n - 1\right)$?}

- Will make "cluster" disappear? No.

Show middle part of CombineCascOnOffData3p5n12.ps

- Then can formulate and address "bias" problems?

- Still need to tackle problems with bias in $\hat{n}$???

Future Research

1. Search for reason behind $\hat{n}$ bias.

2. Estimate of n other than variance matching – e.g. quantile

3. Consider alternative parametrization.

4. Bias is originated in the estimation of $\hat{t}_{on}$ and $\hat{t}_{off}$.

    - Properties of Estimates: unbiased? …

    - How does it affect the parameter estimates?